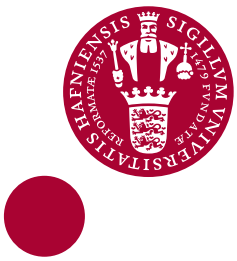

Correlations and Non-Linear Probability Models

Breen, Richard; Holm, Anders; Karlson, Kristian Bernt

Published in:
Sociological Methods & Research

DOI:
10.1177/0049124114544224

Publication date:
2014

Document version
Peer reviewed version

Document license:
Unspecified

Citation for published version (APA):
Breen, R., Holm, A., & Karlson, K. B. (2014). Correlations and Non-Linear Probability Models. *Sociological Methods & Research*, *43*(4), 571-605. https://doi.org/10.1177/0049124114544224

Correlations and Non-Linear Probability Models

Richard Breen*, Anders Holm**, and Kristian Bernt Karlson***

THIS PAPER IS PUBLISHED IN
SOCIOLOGICAL METHODS & RESEARCH, 2014, VOL. 43, NO. 4, 571-605
Link: http://smr.sagepub.com/content/43/4/571
This is a post-print (i.e. final draft post-refereeing) version according to SHERPA/ROMEO

* Center for Research on Inequality and the Life Course, Department of Sociology, Yale University, email: richard.breen@yale.edu.

** Department of Sociology, University of Copenhagen; SFI –The Danish National Centre for Social Research, email: ah@soc.ku.dk.

*** SFI - The Danish National Centre for Social Research; Danish School of Education, Aarhus University, email: kbk@dpu.dk.

Tuesday, May 28, 2013

Acknowledgements: We would like to thank the SMR editor and reviewers and the participants at the RC28 conference held in Trento, Italy, in May 2013, for helpful comments on this paper.

Correlations and Non-Linear Probability Models

Abstract

Although the parameters of logit and probit and other non-linear probability models are often explained and interpreted in relation to the regression coefficients of an underlying linear latent variable model, we argue that they may also be usefully interpreted in terms of the correlations between the dependent variable of the latent variable model and its predictor variables. We show how this correlation can be derived from the parameters of non-linear probability models, develop tests for the statistical significance of the derived correlation, and illustrate its usefulness in two applications. Under certain circumstances, which we explain, the derived correlation provides a way of overcoming the problems inherent in cross-sample comparisons of the parameters of non-linear probability models.

Correlations and Non-Linear Probability Models

Both textbook expositions and discussions of research findings commonly interpret the coefficients of non-linear probability models (henceforth NLPMs) such as logits, probits, the ordered logit and probit, and the multinomial logit, as the coefficients of an underlying latent linear model, albeit identified only up to scale. In this paper—drawing on and extending the work of McKelvey and Zavoina (1975)—we point out that the coefficients of NLPMs are, in fact, closely related to the correlations between the dependent and predictor variables of the latent linear model, and, in some circumstances, interpreting them in the light of this could be particularly useful. This is especially so when we want to compare the effects of the same variable in the same NLPM fitted to different groups (for the problems in doing this see Allison 1999).

We first develop the logit and probit models in a latent variable framework and show how their coefficients can be used to recover the correlation between a predictor variable, x, and y^*, the underlying latent dependent variable. We discuss the conditions under which the correlation coefficient is likely to prove useful and we provide statistical tests of the derived correlation coefficient and for differences in correlations between samples. We also show how to derive the correlation in the ordered and multinomial case. We conclude the paper with two examples dealing with trends in educational inequality. The first is a reanalysis of data on educational inequality in Europe (Breen *et al* 2009). The second uses GSS data to study the trend in educational inequality in the U.S. among cohorts born between 1930 and 1969. Four appendices contain some technical details of our approach.

A Latent Variable Formulation of Logit and Probit Models

Let *y** denote a continuous outcome variable, and let *x* be a predictor variable whose effect on *y** we want to estimate. We can write a regression model for *y** as

$$y^* = \alpha + \beta_{y^*x} x + u, \text{ with } sd(u) = \sigma_u. \quad (1)$$

Here u is a random error term and σ_u its standard deviation, which summarizes that part of the variation in *y** unexplained by *x*. α is the intercept of the model, and β_{y^*x} captures the effect of *x* on *y**. α, β_{y^*x}, and σ_u are all unknown parameters.

However, we assume that *y** is unobserved and instead we observe:

$$\begin{aligned} y &= 1 \text{ if } y^* > \tau \\ y &= 0 \text{ if } y^* \le \tau. \end{aligned} \quad (2)$$

This means that we only observe whether an observation's value of *y** is greater or less than a constant, τ, which is a threshold parameter whose value is unknown. Following a standard latent variable formulation of discrete choice models we rewrite the error term in (1) such that $u = s\omega$. For the logit model, we assume that ω is a standard logistic random variable, with mean zero and variance $\pi^2/3$ and *s* is a scale parameter, yielding a variance of $\sigma_u^2 = s^2\pi^2/3$ for the error term in (1). For the probit model, we assume that ω is a standard normal random variable, with mean zero and unit variance and *s* is a scale parameter, yielding a variance of $\sigma_u^2 = s^2$ for the error term in (1).

For the logit case we specify the probability of success as a function of *x*:

$$\Pr(y=1) = \Pr(y^* > \tau) = \Pr\left(\frac{u}{s} > -\left[\frac{\alpha-\tau}{s} + \frac{\beta_{y^*x}}{s}x\right]\right) = \frac{\exp\left(\frac{\alpha-\tau}{s} + \frac{\beta_{y^*x}}{s}x\right)}{1+\exp\left(\frac{\alpha-\tau}{s} + \frac{\beta_{y^*x}}{s}x\right)}, \quad (3a)$$

where the final equality holds because we assumed ω to be a standard logistic random variable. Taking the logarithm of the odds of the probability in (3a), we obtain the well-known logistic regression model:

$$\text{logit}(\Pr(y=1)) = \frac{\alpha-\tau}{s} + \frac{\beta_{y^*x}}{s}x = a + b_{yx}x\,. \tag{3b}$$

For the probit case we also specify the probability of success as a function of *x*:

$$\Pr(y^* > \tau) = \Phi\left[\frac{\alpha-\tau}{s} + \frac{\beta_{y^*x}}{s}x\right] = \Phi(a + b_{yx}x) \tag{4}$$

Where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, and where the first equality holds because we assumed ω to be a standard normal random variable.

The intercept of the logit or probit model depends on the intercept of the underlying linear model (α) and the threshold parameter (τ) and these two cannot be recovered separately, being absorbed in the intercept, *a*, of the logit or probit. The parameters b_{yx} in the logit and probit models are equal to the regression coefficients from the underlying linear model in (1) divided by the scale parameter, which is a function of the underlying conditional error variance. Thus in probit and logit models we can identify regression coefficients only up to scale, which is a function of the conditional standard deviation of the latent outcome variable. This presents particular difficulties for parameter comparisons, whether these are between parameter estimates for the same variable in different model specifications (for example, with and without control variables; see Karlson, Holm and Breen 2012) or parameter estimates for the same variable in the same models fitted to different samples (Allison 1999). In both cases, real differences in a variable's effect will be confounded by possible differences in scaling.

The issues pertaining to scale identification of coefficients from non-linear probability models have wide-ranging consequences for comparative research analyzing discrete outcomes.

Most significantly, because these coefficients are inherently standardized, researchers cannot generally follow the recognized advice of using "unstandardized coefficients" in group comparisons (Tukey 1954; Blalock 1967). Consequently, we need to look to other metrics that might prove useful in certain areas of sociological research.

The Problem in Group Comparisons When Not Assuming a Latent Outcome Variable

The previous derivations hold under the assumption that we observe the latent outcome, y^*, via its binary manifestation, y. However, in some sociological applications, motivating the logit model in terms of latent variables is less obvious. Yet, even when the binary outcome variable can be said to be truly binary, the problem in comparing logit or other nonlinear probability model coefficients across groups remains. Using the distinction between marginal and conditional models (Agresti 2002), in Appendix A we show that the group comparison problem also applies when the outcome can be said to be truly binary and the latent variable formulation consequently does not apply. In either case, group comparisons of coefficients are distorted by differences in unmeasured heterogeneity across groups even when the unmeasured heterogeneity is independent of the observed predictor variables.

The Correlation and Its Relation to the Logit or Probit Coefficient

Rather than interpreting logit and probit coefficients as underlying effects identified up to scale, they can also be interpreted as functions of the correlation coefficient between the predictor x and the underlying latent variable, y^*. To show this, we use results from the early literature on the relationship between coefficients from linear models and correlations (e.g., Blalock 1964, 1967; Linn and Werts 1969; Theil 1972). The only difference is that we apply the results to y^*, allowing us to derive the correlation coefficient from the probit and logit model.

We can write the variance of the underlying latent variable, *y**, as

$$\text{var}(y^*) = \beta_{y^*x}^2 \text{ var}(x) + \text{var}(y^* \mid x) = \beta_{y^*x}^2 \text{ var}(x) + s^2 \text{ var}(\omega) \tag{5}$$

and the correlation between *y** and *x* is equal to

$$r_{y^*x} = \beta_{y^*x} \frac{sd(x)}{sd(y^*)}.$$

Using (5) and the fact that the logit or probit coefficient, b_{yx}, equals $\frac{\beta_{y^*x}}{s}$ we can write

$$r_{y^*x} = \frac{b_{yx} s \cdot sd(x)}{\sqrt{b_{yx}^2 s^2 \text{ var}(x) + s^2 \text{ var}(\omega)}} = \frac{b_{yx} \cdot sd(x)}{\sqrt{b_{yx}^2 \text{ var}(x) + \text{var}(\omega)}}, \tag{6}$$

since the *s* terms cancel.

For the logit we substitute $\text{var}(\omega) = \pi^2/3$ and for the probit $\text{var}(\omega) = 1$. With a single predictor variable the correlation in (6) equals the fully standardized coefficient suggested by McKelvey and Zavoina (1975: 115).

Using (6) we can write the logit or probit coefficient as a function of the correlation coefficient. To do so, we solve for *k* in $b_{yx} = k \cdot r_{y^*x}$. Using that $s = \frac{sd(y^* \mid x)}{sd(\omega)}$ and writing the correlation and b_{yx} coefficient in terms of variances and covariances,

$$\frac{\text{cov}(x, y^*)}{\text{var}(x) s} = k \frac{\text{cov}(x, y^*)}{sd(x) sd(y^*)},$$

we obtain

$$k = \frac{\text{cov}(x, y^*) sd(\omega) sd(x) sd(y^*)}{\text{cov}(x, y^*) \text{ var}(x) sd(y^* \mid x)} = \frac{sd(\omega) sd(y^*)}{sd(x) sd(y^* \mid x)} = \frac{1}{\sqrt{1 - r_{y^*x}{}^2}} \frac{sd(\omega)}{sd(x)}$$

and thus

$$b_{yx} = \frac{r_{y^*x}}{\sqrt{1-r_{y^*x}^2}} \frac{sd(\omega)}{sd(x)} \quad . \tag{7}$$

Thus logit and probit coefficients can be expressed as the square root of the ratio of the explained to unexplained variance of *y** scaled by the ratio of the standardized standard deviation to the standard deviation of the predictor variable.[1] From this we see that a logit or probit coefficient has two parts, one which is scale invariant, $\frac{r_{y^*x}}{\sqrt{1-r_{y^*x}^2}}$, and one which is scale-dependent, $\frac{sd(\omega)}{sd(x)}$. As we will argue, this separation provides an entry into addressing the problem of making comparisons between groups when using NLPMs.

The derivative of the logit or probit coefficient with respect to the correlation is

$$\frac{\partial b_{yx}}{\partial r_{y^*x}} = \frac{sd(\omega)}{sd(x)(1-r^2)^{3/2}}$$

The derivative tells us that the logit or probit coefficient increases as the correlation deviates from zero.

From the forgoing we see that the relationship between the correlation and the b_{yx} coefficients depends only on known quantities—namely $sd(\omega)$, which is known (by assumption), and $sd(x)$ which is known from the data. This is in contrast to the relationship between b_{yx} and β_{y^*x} which depends on the unknown quantity, *s*.

[1] In the probit case, $sd(\omega) = 1$, and if we standardize *x* to have unit standard deviation, the squared probit coefficient will equal the ratio of explained to unexplained variance in the underlying latent variable regression.

Partial, Semi-Partial, and Fully Standardized Partial Coefficients

Equation (6) applies when there is a single predictor variable. With two or more predictor variables, we can derive the partial correlation of x and y^* given z (this is demonstrated in Appendix B) as

$$r_{y^*x.z} = \frac{b_{yx.z}\, sd(x \mid z)}{\sqrt{b_{yx.z}^2 \operatorname{var}(x \mid z) + \operatorname{var}(\omega)}}, \tag{8}$$

and, as a consequence, we can write the partial logit or probit coefficient as

$$b_{yx.z} = \frac{r_{y^*x.z}}{\sqrt{1 - r_{y^*x.z}^2}} \frac{sd(\omega)}{sd(x \mid z)}. \tag{9}$$

The extension to the case with several control variables is straightforward. Despite this, the extension from the simple to the partial case has never before been demonstrated, as far as we are aware. As in the linear case (Blalock 1967: 133), the partial correlation in (8) will generally differ from the fully standardized partial coefficient, $b_{yx.z}^*$, used by McKelvey and Zavoina (1975),

$$b_{yx.z}^* = \frac{b_{yx.z}\, sd(x)}{\sqrt{b_{yx.z}^2 \operatorname{var}(x) + b_{yz.x}^2 \operatorname{var}(z) + 2 b_{yx.z} b_{yz.x} \operatorname{cov}(x, z) + \operatorname{var}(\omega)}}, \tag{10}$$

their analytical relation being:

$$b_{yx.z}^* = r_{y^*x.z} \frac{\sqrt{1 - r_{y^*z}^2}}{\sqrt{1 - r_{xz}^2}},$$

Similarly, we derive the semi-partial[2] correlation between y^* and x given z as

[2] The semi-partial correlation is also known as the part correlation.

$$r_{y^*(x.z)} = \frac{b_{yx.z} sd(x \mid z)}{\sqrt{b_{yx.z}^2 \operatorname{var}(x) + b_{yz.x}^2 \operatorname{var}(z) + 2 b_{yx.z} b_{yz.x} \operatorname{cov}(x, z) + \operatorname{var}(\omega)}}.$$

meaning that the semi-partial correlation and the fully standardized partial coefficient have the following relation:

$$b_{yx.z}^* = r_{y^*(x.z)} \frac{sd(x)}{sd(x \mid z)} = r_{y^*(x.z)} \frac{1}{\sqrt{1 - r_{xz}^2}}.$$

Thus, the fully standardized partial coefficient can be viewed as a rescaled version of the semi-partial correlation, with the scale factor being the square root of the proportion of the variance in *x* that is unexplained by *z*, $\sqrt{1 - r_{xz}^2}$.

When Is the Correlation a Useful Metric?

Using correlations rather than logit or probit coefficients themselves implies a shift of focus for researchers interested in group comparisons. In this section, we clarify the conditions under which the correlation coefficient is a useful metric for group comparisons and suggest which methods researchers might consider if the conditions we outline are not met.

When the Correlation Is Useful

We have shown how to derive the correlation from the parameters of an NLPM: but when, and why, should we want to do this? We outline three circumstances.The first circumstance in which the correlation will be useful is when we want to compare the variation in *y** between and within values of *x*. This is particularly the case when *x* is categorical, as, for example, in studies that focus on the relationship between educational attainment and social class background. In many studies the observed dependent variable, *y*, is whether or not a student makes a given educational transition (such as the transition from High School to College) and the unobserved *y** could be

interpreted as his or her propensity to do so. In these sorts of analysis we are often interested in the changing relationship, over birth cohorts, between dummy variables, *x*, representing social classes, and *y** (e.g. Shavit and Blossfeld 1993). The conventional approach would take comparisons of β_{y^*x} over cohorts as the ideal measure, and one would try to recover these by estimating the corresponding logit or probit coefficients, b_{yx}. Logit or probit coefficients declining in magnitude over successive birth cohorts would then suggest equalization in the particular educational transition. But the problem with this interpretation is that we cannot know the extent to which declines over cohorts in b_{yx} are due to declines in β_{y^*x} or to increases in the residual variation of *y**.

An alternative measure of equalization is the degree to which variation in the propensity to make the transition exists between students from different classes, relative to the variation that exists among students within the same class. A decline over cohorts in the variation in *y** tells us that there is less variation as a whole: in this case the correlation would decline only if the variation between students from different classes was declining more quickly than the total variation. Thus, if the correlation was constant over birth cohorts, b_{yx} could decline only as a result of a reduction in the variance of *y** and/or an increase in the variance of *x*. But it was to remove these 'spurious' influences on the measure of inequality that led sociologists to the use of NLPMs in the first place (Mare 1981): by the same argument, they should then prefer the (partial) correlation as a measure of educational inequality.

The second circumstance in which the correlation can be informative is when we care about the relative, rather than the absolute, value of the outcome variable. In studies of intergenerational income mobility, for example, it is reasonable to focus on an individual's position within a given income distribution (that is, relative to others) rather than, or in addition

to, the individual's absolute level of income. For this reason, the correlation of (the logarithm of) parent's and child's incomes is sometimes preferred, as a measure of intergenerational immobility, to the elasticity (the coefficient from the regression of log of child's income on log of parent's income) because it is insensitive to differences in the variance of income in the parental and child generations (see Björklund and Jäntti 2009).

Because we can recover the correlation between y^* and one or more predictor variables, we can estimate the standardized regression coefficients linking them, and these will be particularly useful when we care about relative rather than absolute position. Consider the following example:

> one would expect the impact of a scholar's rate of publication on his or her academic salary to be affected by the mean and variance of each of these two variables within the scholar's field. The publication of two papers per year more than the mean publication rate in mathematics is a more impressive performance than the same achievement in chemistry because the variance of publication rates is much larger in the latter field ... Since the variance of annual salaries also differs across fields, it is unreasonable to expect that publishing one additional paper during a period of time will have an equivalent impact on annual salary across fields. In contrast, it seems more reasonable to expect that the impact of increasing one's publication rate by one field-specific standard deviation will have an equivalent impact on field-specific standard scores for academic salary across fields (Hargens 1976: 252).

If we substitute the phrase 'propensity to obtain tenure' for 'annual salary' we have a binary outcome and here an interpretation of logit or probit coefficients in terms of the x-y^* correlations or standardized regression coefficients should be preferred over an interpretation of them as underlying regression coefficients identified up to scale.

The third circumstance concerns the situation in which researchers are interested in conditional inference; that is, interested in the association between two variables (y^* and x) net of a third variable (z). For example, stratification researchers might want to compare across birth cohorts the magnitude of the association between college completion and parental social status net of race. In such situations, the semi-partial correlation will often be preferred over the partial correlation. The latter correlation partials out the variance in both y^* and x explained by z, while the former only partials out the variance in x explained by z. In other words, the semi-partial correlation (or its squared counterpart) is a measure of the fraction of explained variance in y^* due to x net of z relative to the unconditional or total variance in y^*. In contrast to the partial correlation, the semi-partial correlation is unaffected by the predictive power of z on y^*, thereby not conflating differences in the predictive power of z in group comparisons.

An alternative to the semi-partial correlation is the fully standardized partial coefficient of McKelvey and Zavoina (1975), stated in Equation (10). While this coefficient is just a rescaled version of the semi-partial correlation, it is useful whenever researchers are concerned with the relative, rather than the absolute, value of the outcome variable, as outlined above. In this situation, the fully standardized partial coefficient can be interpreted as the expected standard deviation change in y* for a standard deviation change in x, given z. For example, stratification researchers might want to know whether the dependency of the son's relative position in the socioeconomic status distribution on the relative position of his father in the father's corresponding distribution, once race is controlled, has changed over the 20th century. In such a situation, the fully standardized partial coefficient might be preferred.

One thing which is clear from these examples of where the correlation may prove useful, is that, when these conditions prevail, interpreting the coefficients of NLPMs in terms of

correlations also provides a solution to the problem of comparing the effects of variables across samples, between which the residual variation, captured in the scale parameter, *s*, might differ.

When the Correlation Is Not Useful

The conditions under which the correlation is a useful metric for group comparisons may not always be met. In this subsection we first briefly outline these conditions and then discuss two other methods that might be useful when the conditions are not met.

Unstandardized or Standardized Coefficients?

Unstandardized regression coefficients have long been preferred over standardized coefficients in social science research because the former are unaffected by the distributions of the variables involved in the model (Blalock 1967; Kim and Mueller 1976; Schoenberg 1972; Tukey 1954). In the linear regression model, the unstandardized regression coefficient of *x* on *y* is given by

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)},$$

yielding the expected change in *y* for a unit change in x. The standardized—or equivalent correlation—coefficient is given by

$$r = \beta \frac{sd(x)}{sd(y)},$$

and is affected by both the effect of *x* on *y* and the distributions of *x* and *y*. Using standardized or correlation coefficients for group comparisons will conflate differences in true effects, captured

by β, with potential differences in the distributions of both *x* and *y*, captured by *sd(x)* and *sd(y)*. For this reason, and for reasons stated throughout this paper, using the correlation coefficient for group comparisons is not informative about differences in effects; it is merely informative about differences in the predictive power of *x* on *y* (Achen 1977; King 1986) or about the association between relative positions in the distributions.

For example, the correlation coefficient would not be useful for comparing the influence of high school grades on college enrollment between men and women. Because the correlation standardizes grades and the latent college enrollment propensity within gender, we cannot know whether gender differences in correlations reflect differences in effects or differences in the dispersion in the latent propensity distribution or in the grade distribution. If the effect, β, and the dispersion in the latent propensity, *sd(y)*, were the same for men and women, but the dispersion in the grade distribution, *sd(x)*, was larger among men than women, then we would see larger correlations among men than women. Such differences would not be informative about gender differences in the influence of grades on college enrollment, even if we interpreted them as associations between relative positions in the distributions. This interpretation would imply that women (men) only compete with other women (men) in obtaining college positions. But this is unlikely to be true in most countries. Similar examples can be given of comparison across subpopulations that compete for the same goods

Nevertheless, although using the correlation for comparative purposes is restricted in many respects, the same applies to nonlinear probability models in which the coefficients, by design, are inherently standardized on the latent outcome variable via the scale parameter, *s*. Since the true effect, β_{y^*x}, cannot be separately identified, obtaining unstandardized coefficients is not possible in nonlinear probability models. And since the scale parameter, *s*, depends on both

the total variance in the latent outcome and the predictive power of *x*, coefficients of nonlinear probability models are difficult to interpret.[3] This is precisely the reason why we propose viewing these coefficients in terms of correlations which, although limited in their use, have a well-defined interpretation.

However, viewing nonlinear probability model coefficients as correlation coefficients is only one way of approaching the limitations of the inherent standardization of these coefficients. As we have already pointed out, the correlation coefficient is standardized on both the predictor variable and the outcome variable. Yet, because we know the scale of the predictor variable, *x*, standardizing the coefficient on *x* is not necessary. Researchers may instead opt for the coefficient standardized on the latent outcome alone. This coefficient was introduced in the seminal work by Winship and Mare (1984).[4] In the simple case with only two variables, it is given by

$$b_{yx}^{YSTD} = \frac{\beta_{y^*x}}{sd(y^*)} = \frac{b_{yx}}{\sqrt{b_{yx}^2 \operatorname{var}(x) + \operatorname{var}(\omega)}}.$$

This partially standardized coefficient differs from the fully standardized counterpart in (6) by not including *sd(x)* in the numerator. It yields the expected standard deviation change in *y** for a unit change in *x*, and can easily be extended to the multiple case,

$$b_{yx.z}^{YSTD} = \frac{\beta_{y^*x.z}}{sd(y^*)} = \frac{b_{yx.z}}{\sqrt{b_{yx.z}^2 \operatorname{var}(x) + b_{yz.x}^2 \operatorname{var}(z) + 2b_{yx.z}b_{yz.x} \operatorname{cov}(x,z) + \operatorname{var}(\omega)}},$$

[3] The interpretation of coefficients of nonlinear probability models is even more complicated when several predictor variables are included. In this situation, the coefficients reflect the true effects of interest and the predictive power of all predictor variables.

[4] Breen and Karlson (2013) discuss the advantages of using coefficients standardized on the outcome in causal inference involving nonlinear probability models and show the equivalence between these coefficients and Cohen's *d*, an effect size metric much used in evidence-based research.

In the multiple case, the coefficient expresses the expected standard deviation change in y^* for a unit change in x, holding z constant.

The partially standardized coefficient may prove useful in certain areas of comparative research. For example, in studying the black and white gap in college attainment over the 20th century U.S, researchers might want to know whether the gap in the underlying propensity to complete college narrowed over time. In this situation, the partially standardized coefficient could be used to compare the black-white gap in y^*, measured in standard deviations, across different time periods. Such an analysis would be informative about the changes in the black-white gap in college completion as a positional good. Moreover, if researchers were interested in comparing these gaps net of socioeconomic status, the partially standardized coefficient for the multiple case could be used.

Given the often complicated interpretation of nonlinear probability model coefficients, the partially standardized coefficient is an attractive alternative that acknowledges the inherent standardization on the latent outcome variable and retains the scale interpretation of the predictor variable of interest. Because this coefficient is not sensitive to the distribution in x it may be useful in certain areas of comparative research. We now turn to two other alternatives to using logit and probit models in group comparisons in sociological research.

Using Predicted Probabilities

The issue of comparing logit or probit coefficients across groups was addressed by Ai and Norton (2003) whose approach uses the predicted values from the logit or probit model, and Long (2009) shows how to implement the method when researchers want to use it to make group

comparisons. Because predicted probabilities are non-linear, group differences in them will vary across the distribution of the predictor variable of interest. This means that an interaction effect cannot be reduced to a single quantity, but has to be studied graphically across the distribution of the predictor variable. To implement this idea, Long suggests using a logit or probit model to estimate, for each group, the predicted probability of the outcome at different values of a predictor variable of interest and then plotting the difference between the groups' predicted probabilities across the distribution of the predictor variable. Long (and Ai and Norton) also suggest statistical tests of group difference in probabilities.

This approach could be applied when hypotheses about group differences can be tested through their implications for probabilities. It is also likely to be valuable in showing that the magnitude of group differences can vary across the distribution of predictor variables. This can often best be seen graphically; indeed, as Greene (2010) notes, such graphical output is needed whenever we are to evaluate interaction effects on the probability margin. Nevertheless, this method has some limitations. As Greene (2010: 295) observes, "partial [marginal] effects are neither coefficients nor elements of the specification of the model. They are implications of the specified and estimated model" (brackets added). One consequence of this property is that tests of significance of differences in predicted probabilities cannot be interpreted in the way we would interpret tests of significance of differences in coefficients. Furthermore, while the method is simple to implement in models with binary outcomes, in the ordinal or multinomial case the graphical output will rapidly become cumbersome.

Using the Heteroskedastic Non-Linear Probability Model

As well as pointing to the problem of making comparisons across groups when using non-linear probability models, Allison (1999) also proposed a solution that involved estimating the ratio of

error variances between groups. Williams (2009) showed that Allison's method is a special case of what is variously termed a "heteroskedastic non-linear probability model", a "heterogeneous choice model", and a "location-scale model". This model is characterized by an equation for the variance as well as the mean. The variance can be written as a function of any relevant predictor variables, but, in the context of group comparisons, we would be most interested in models that allowed the variance to differ according to dummy variables for groups. For example, applying the model to the male-female biochemist data originally used by Allison, Williams (2009: 540) reports that the residual variance for women is 1.35 times larger than for men.

What appears to have been overlooked in discussions of these models, however, is the issue of identification; that is, our ability to draw inferences about the parameters of interest, such as differences in true coefficients across groups. If we consider the case of comparisons using the ordered logit or probit model (see equation (11), below)[5], then two groups can differ in one or all of their threshold parameters, τ_j, their regression coefficients, β, and their scaling factor, s. A model in which all of these differ is not identified. This follows because such a model is equivalent to fitting a separate model for each group and so, if it were identified, this would allow us to recover the group specific scaling factor from that group's ordered logit model—something which we know is not possible. Thus, to identify the relative sizes of the different groups' scaling factors, it is necessary to impose a group constraint on either or both of τ_j and β (see Williams 2009: 551, whose application of the heterogeneous choice ordered probit model is "contingent on the thresholds being the same for both men and women").

[5] Everything that follows holds for binomial as well as ordinal outcomes and, indeed, for non-linear probability models as a whole.

Whenever we have reason to believe that a constraint on β or τ_j is defensible, the model may be a good choice for making comparisons (see, e.g., Mouw and Sobel 2001) because we can recover the group difference in the true coefficients of interest. However, as Long (2009) argues, in sociology we rarely have either the body of accumulated knowledge or strong theory that would make the imposition of such constraints anything other than ad hoc.[6] Thus, whenever the condition cannot be met, differences in unknown thresholds will be confounded with differences in unknown scaling factors, and so heteroskedastic non-linear probability models will not solve the group comparability problem.

Some Further Results

Extension to the Ordered and Multinomial Case

In the ordered logit and probit we observe not merely whether *y** exceeds an unknown value or not (as in the binary logit or probit) but into which interval of *y** a given observation falls. Thresholds divide the range of *y** into disjoint and exhaustive intervals, $\tau_1 < \tau_2 < ... < \tau_J$ where $\tau_1 = -\infty$ and $\tau_J = \infty$. These allow us to define $y^O = j$ if $\tau_{j-1} < y^* < \tau_j$ where y^O is an ordered, discrete variable with *J* categories and the ordered probit or logit model takes the form:

$$\mathrm{g}(\Pr(y^* > \tau_j)) = \frac{\tau_j}{s} + \frac{\beta}{s} x_t, \qquad (11)$$

where *g(.)* is a link function—the logit or the cumulative normal in the ordered logit and ordered probit respectively—and $b = \beta/s$ is the ordered logit or ordered probit coefficient of *x*. Because the underlying latent variable model in the ordered models is the same as in the binary case (the

[6] It is, of course, possible to make the residual variance depend on other variables. But in so far as these variables do not wholly determine the error variance we are still in the dark as to how much it differs between groups.

difference between them being only in the degree to which y^* is observed), the expressions we have derived for the correlation and partial correlation (and for the coefficient of determination which we derive in Appendix D) apply directly to the ordered case.

The multinomial case is more complicated because we have a number of underlying latent variables. We define x as before, but we now define y^*_a as the propensity of an individual to choose alternative a from among a set of A possibilities. We assume that x and y^*_a are related such that

$$y_a^* = \alpha_a + \beta_a x + s_a \cdot u_a, \tag{12}$$

where β_a captures the effect of x for alternative a, u_a is an alternative-specific random error term, which is independent across alternatives and follows a standard type-I extreme value distribution, and s_a is an alternative-specific scale parameter. We only observe which of the A alternatives the individual actually chooses and we assume that the individual chooses the alternative for which he or she has the greatest propensity:

$$y = a \text{ if } y_a^* - y_{a'}^* > 0, \ \forall\, a' \neq a$$

McFadden (1974) shows that the probability of choosing alternative a given x is equal to

$$\Pr(y = a) = \frac{\exp(k_a + b_a x)}{\sum_a \exp(k_a + b_a x)}, \tag{13}$$

with normalization such as $k_{a=1} = b_{a=1} = 0$ to secure identification, and where

$$k_a = \frac{\alpha_a}{s_a} \text{ and } b_a = \frac{\beta_a}{s_a}.$$

The odds of choosing a rather than the baseline category a' are $\exp(k_a + b_a x)$ and from this we can derive the correlation between $y_a^* - y_{a'}^*$ and x in much the same way as in the binary case, when we add one important qualifier: The standard deviation of x used in the correlation should

be calculated not on the entire sample, but on the sample which has chosen either the alternative, *a*, or the reference category, *a'*. The reason for this is that the standard deviation of *x* on the contrast-specific sample can differ from the standard deviation of x in the full sample. We may then derive the correlation as

$$r_{(y_a^*-y_{a'}^*),x} = \frac{b_a sd(x)_a}{\sqrt{b_a^2 \operatorname{var}(x)_a + \pi^2/3}}, \tag{14}$$

where we define $sd(x)_a$ as the standard deviation of *x* on the sample pertaining to the alternative *a* versus the references contrast *a'* (and $\operatorname{var}(x)_a$ is its squared counterpart). The expression in (14) generalizes easily to any contrast between two alternatives.

Sensitivity to Mis-Specified Error Term

As we saw earlier, the correlation between a predictor and the latent outcome can be recovered from the parameters of a NLPM because we know the variance of the predictor and $\operatorname{var}(\omega)$ is given by assumption. Typically, when *y* is a binary realization of *y** we assume ω to have either a standard normal (for the probit) or standard logistic (for the logit) distribution. But this raises the question: how sensitive are our estimates of the correlation to such an assumption? To answer it we ran a Monte Carlo simulation in which we fitted logit and probit models to data in which we varied the true underlying error distribution, the sample size, and the magnitude of the correlation. In Table 1 we report the mean, over 500 replications, of the absolute deviation from the true correlation. It is noticeable that, except for the uniformly distributed error, the metric is largely insensitive to pure misspecifications, with the mean deviations for the misspecified cases differing little from those where the model is correctly specified (the logit-logistic and probit-

normal cases). This holds even for a lognormal distribution with a skewness of -1 where the error term is highly asymmetric. As we would have expected, the sensitivity is less pronounced the larger the sample size.

--TABLE 1 HERE --

Relationship to the Polyserial Correlation Coefficient

The correlation coefficient we present in this paper is closely related to the polyserial correlation coefficient widely used in psychometrics (Cox 1974; Olsson, Drasgow, and Dorens 1982). However, we believe that our approach is more general, more flexible, and easier to implement. The polyserial correlation assumes that y^* and x follow a bivariate normal distribution, whereas our measure places no assumptions on the distribution of x and allows y^* to have distributions other than normal (e.g., the logistic, complementary log-log). Our method is easily generalizable to partial correlations (using partial logits and conditional standard deviation of x) whereas the partial polyserial correlation has to be derived from several polyserial correlations, severely complicating the derivation of standard errors. Furthermore, unlike the polyserial coefficient, our method extends to partial correlations without making any assumptions about the distribution of the control variable, z.

Significance Testing

Because the correlation is asymmetrically distributed, tests of its statistical significance are usually undertaken on the Fisher transformed coefficient. The Fisher transformation of the correlation coefficient, r, is

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \text{arctanh}(r) \tag{15}$$

F(r) is approximately normally distributed with a known mean and standard error under the null. However, in our case the correlation is not directly calculated from data: rather it is computed as a function of an estimated quantity, b_{yx}. In calculating the standard error of *F(r)* we therefore take this into account, but, having done this, we can then exploit the normality of *F(r)* to calculate confidence intervals and significance tests.

We compute the standard error of *F(r)* using the delta method. The asymptotic standard error obtained from this approach depends on the partial derivative of *F(r)* with respect to *b* and we compute this using the chain rule $\frac{\partial F(r)}{\partial b_{yx}} = \frac{\partial F(r)}{\partial r}\frac{\partial r}{\partial b_{yx}}$. The standard error also depends on the variance of b_{yx}. We can write the asymptotic variance of *F(r)* as:

$$\left[\frac{dF(r)}{dr}\frac{dr}{db_{yx}}\right]^2 \text{var}(b_{yx}) \tag{16}$$

where

$$\frac{dF(r)}{dr}\frac{dr}{db_{yx}} = \frac{sd(x)}{\sqrt{b_{yx}^2\,\text{var}(x) + \text{var}(\omega)}}$$

The asymptotic standard error of *F(r)* is given by the square root of equation (16). Because the Fisher transformed correlation is asymptotically normally distributed under the null, we can test the null hypothesis of zero correlation via:

$$Z = \frac{F(r)}{se\big(F(r)\big)}. \tag{17}$$

We analyzed the power of the test score in (17) using a Monte Carlo study. The results are reported in Table 2. We generated the data using equation (1) with *x* normally distributed and *u* having a standard logistic distribution and we derived the estimated correlation from a logit

model. We varied the true correlation and the sample size and, for each scenario, we report the percentage of times out of 1000 replications that our test rejected the null hypothesis (based on a .05 critical value). In scenario A of Table 2 the null hypothesis is true, and here our test rejects the null hypothesis 5% of the time or less. In scenarios B through F, the null hypothesis is false, and our test fails to reject it only when the sample size is small and the true correlation is quite small. Even in samples of 200 and for a correlation of 0.25 a false null has a 90% rejection rate. In samples of 1,000 observations or larger, we obtain a power of above 80 percent even for small correlations.

-- TABLE 2 HERE --

The Fisher transform can easily be extended to test the null hypothesis of zero difference between groups in their correlations:

$$Z = \frac{F(r_{y^*x,2}) - F(r_{y^*x,1})}{\sqrt{\text{var}\left(F(r_{y^*x,1})\right) + \text{var}\left(F(r_{y^*x,2})\right)}} \tag{18}$$

Here the subscript 1 and 2 indicates different samples between which we want to compare correlations. We conducted a Monte Carlo study to evaluate the power of this test. The results can be found in Appendix C. They show that the larger the difference between correlations and the larger the sample size, the more likely the test is to reject the null hypothesis of no difference in correlations. With a sample of 1000 the test is effective at rejecting the null hypothesis when the difference in correlations is 0.2 or greater: for a sample of 5000 the test performs well for differences in correlations of more than about .07.

Applications

The final part of our exposition applies the methods we have proposed to two examples that illustrate the relationship between the correlation and the coefficients of NLPMs. Our first example is a reanalysis of data on trends in educational inequality in Europe (Breen *et al* 2009). Using the correlation slightly modifies the findings of the original study and we explain why by investigating the source of the differences that we find between the ordered logit coefficients on which the original conclusions were based and the correlations on which we rest our conclusions. Our second example uses data from the General Social Survey (GSS) to examine trends in the relationship between educational attainment and father's socio-economic index over cohorts born between 1930 and 1969, and here we find that the differences between trends based on an ordered logit model and those based on the correlation are much more pronounced.

Non-Persistent Inequality in Educational Attainment

Shavit and Blossfeld (1993) summarize the results of their seminal study of inequalities in educational attainment across different cohorts under the title *Persistent Inequality*. They claim that, despite dramatic educational expansion during the twentieth century, all but two (Sweden and the Netherlands) of the thirteen countries studied in their project "exhibit stability of socio-economic inequalities of educational opportunities" (Shavit and Blossfeld 1993: 22). Recently, Breen, Luijkx, Müller and Pollak (2009) have challenged Shavit and Blossfeld's conclusions using data on educational inequality in the twentieth century in eight European countries (see also Breen and Jonsson 2005; Breen, Luijkx, Müller and Pollak 2010).

All these studies base their comparisons between birth cohorts on the use of NLPMs: logits in the case of the Shavit and Blossfeld (1993) volume, the ordered logit in the Breen *et al* (2009, 2010) analysis. This implies that their conclusions may confound real differences in educational

inequality across cohorts with differences in residual variation. Here we re-analyze data from three of the eight countries in the Breen *et al* study and compare the coefficients—and thus the conclusions drawn—from the original model with the corresponding partial correlations. The correlation, as a standardized measure, is particular appropriate in this case because, from the perspective of the study of inequality, education can be seen as a positional good: what matters is a person's position in the distribution rather than his or her absolute level of education. Furthermore, when, as here, the predictor variables are dummies, the correlation has an attractive interpretation, telling us how much of the variance in the latent y^* lies between classes rather than within them.

The data we use relate to men in Germany, France, and the Netherlands, born in one of five cohorts: 1908-24, 1925-34, 1935-44, 1945-54, and 1955-64. The dependent variable is highest level of educational attainment measured using five ordered categories:

1 Compulsory education with or without elementary vocational education,

2 Secondary intermediate education, vocational or general,

3 Full secondary education,

4 Lower tertiary education,

5 Higher tertiary.

The predictor variable is social class origins, based on the respondent's report of his father's occupation when he was growing up. Seven classes are distinguished using the EGP class schema (Erikson and Goldthorpe 1992, chapter 2):

I Upper service,

II Lower service,

IIIa Higher grade routine non-manual workers,

IVab Self-employed and small employers,

IVc Farmers,

V+VI Skilled manual workers, technicians and supervisors,

VIIab+IIIb Semi- and unskilled manual, agricultural, and lower grade routine non-manual workers.

Sample sizes are 17,124 for Germany, 51,705 for France, and 19,751 for the Netherlands. Further detail on the data and on the surveys from which they were obtained can be found in Breen *et al* (2009: 1480-5). These three countries were chosen for reanalysis here because Breen *et al* found significant decline in class inequality in educational outcomes over successive birth cohorts in all three countries, whereas in *Persistent Inequality* a decline was found for the Netherlands but not for Germany (France was not included). The question we now seek to answer is whether Breen *et al's* claim of 'non-persistent' inequality is robust if we use the correlation as our measure of inequality in educational attainment.

We begin our analysis by replicating theirs. Breen *et al* use an ordered logit model to regress the five educational categories, considered an ordinal ranking of educational attainment, on dummy variables representing the origin classes. They fit a separate model to each of the five birth cohorts and present their results in a series of figures, of which the most important is Figure 4 on page 1495 of their paper. Our replication of their analysis is shown here in Figure 1 and this is identical with their Figure 4.

-- FIGURE 1 HERE --

Class I serves as the omitted category, having an implicit coefficient of zero, and the lines plotting the coefficients for the other classes (which measure the extent to which the particular class's log-odds of exceeding any given level of education differ from those of class I) show a

general trend of convergence towards zero as we move from later- to earlier-born cohorts. This trend extends across the whole of the 20th century in the Netherlands but does not begin until after the cohort born 1925-34 in Germany and France. The trend is most pronounced among the initially most disadvantaged classes, particularly farmers (IVc) and unskilled workers (VII+IIIb) (see Breen *et al* 2009: 1494-6).

Figure 2 shows, for each class, the partial correlation (that is, controlling for the effect of the other class dummies) with latent educational attainment, y^*. These are all negative (because they are all measured relative to class I) and so a line in Figure 2 that rises towards zero over cohorts (a notable example is the line for class IVab in the Netherlands) tells us that a declining share of the conditional variance in y* lies between that origin class and class I.

-- FIGURE 2 HERE --

The trend towards greater equality, measured as the convergence of class coefficients, is somewhat less pronounced in Figure 2 (correlations) than in Figure 1 (logits), especially for France, and, in both France and Germany, Figure 2 suggests that the trend to equalization began later (after the 1935-44 cohort) than Breen *et al* found. Trends for particular classes are not always the same in the two figures: e.g. in France the partial correlations show that the positions of classes III and V+VI worsened relative to class I, whereas the logit coefficients in Figure 1 show that their position either remained unchanged or improved. Equation (8) tells us that such discrepancies must be due to changes in the conditional (on the other classes) variance of the

class dummies which will, to a considerable extent, reflect the changing relative sizes of the social classes.[7]

We can investigate the differences between the logit coefficients and the partial correlations by returning to equation (9) that shows that the partial logit coefficient is equal to the square root of the ratio of the explained to the unexplained conditional variation in the latent *y**, $\frac{r_{y^*x.z}}{\sqrt{1-r^2_{y^*x.z}}}$, scaled by the ratio of the standardized standard deviation of the error ($\pi/\sqrt{3}$ in the logit case) to the residual standard deviation of *x*, $sd(\omega)/sd(x\,|\,z)$. We can therefore decompose the logit coefficients into these two parts. In Table 3 we report these two parts for the three countries. In these expressions, *z* stands for the dummy variables for the classes other than the one being studied. The logit coefficients shown in Figure 1 are the product of these two components.

--TABLE 3 HERE --

The trend in the French results reported in Figure 2 are very similar to those of the ratio $r_{y^*x.z}/\sqrt{1-r^2_{y^*x.z}}$ shown in Table 3. But the more pronounced decline in class origin effects in France seen in the logits of Figure 1 is due to the steep reduction in the ratio $sd(\omega)/sd(x\,|\,z)$. And since $sd(\omega)$ is constant over cohorts, this change reflects the increase in the conditional standard deviations of the dummy variables over cohorts. The Dutch case is similar: increases in the conditional standard deviations of the dummy variables together with declines in their explanatory power give rise to the even more marked equalization shown in Figure 1. But the

[7] The conditional variance of a class dummy is the variance of the residual from a regression of that dummy variable on the dummy variables for the other classes. Thus, change over cohorts in the conditional variance of a particular dummy reflects not just changes in the relative size of that particular class; it is also sensitive to changes in the entire class distribution, albeit in a complicated way.

German experience is somewhat different: here Figures 1 and 2 are more similar because less of the decline in the logits is attributable to change in the conditional standard deviations of the class dummies. As the trend in the ratio, $sd(\omega)/sd(x|z)$, in Table 3 shows, although the conditional standard deviations of the class dummies have mainly increased (and so the ratio reported in the figure declines) they have remained within a narrower range.

In general, changes in the residual standard deviation of the predictor variables could cause logits (or, generally, NLPMs) to exaggerate or underplay trends or differences in the degree to which the predictors account for variation in the latent *y**. If $sd(x|z)$ increases over cohorts, as here, it will exaggerate a declining trend in educational inequality when measured with logit coefficients but underestimate the opposite trend when measured with partial correlation coefficients. A declining $sd(x|z)$ will magnify an increasing trend in logit coefficients but dampen a declining one.

In the case at hand, where changes in $sd(x|z)$ magnify a declining trend, our overall conclusion would nevertheless be that inequalities in educational attainment declined over the 20th century. To support this conclusion, we report in Figure 3 the estimated R^2 values (derived in Appendix D) for each cohort. We find a clear decline in the degree to which class origins account for variation in latent educational attainment. And, because correlations are standardized measures, we can also compare them across countries. In the oldest cohort class origins explained most variation in educational attainment in Germany (30%), but this is the country in which there has been the greatest absolute decline in R^2 and by the youngest cohort class origins account for about 15% of the variance in both Germany and France. In all cohorts the Netherlands displays the lowest amount of variance explained by class origins: in the youngest

cohort this is only 8%. In all three countries the share of the variance in educational attainment lying between classes has been approximately halved over the 20th century: i.e., class origins were a much weaker factor in explaining the variation in educational attainment among men born in the middle of the century than among those born at its start.

--FIGURE 3 HERE—

Trends in Inequality of Educational Attainment in the U.S.

In this example we use data from the General Social Survey (GSS), 1988-2010, to study trends in inequalities in educational attainment across four cohorts born between 1930 and 1969 in the U.S. Hout *et al* (1993) analyzed earlier GSS data using logit models and found no evidence for declining effects of father's occupational prestige across cohorts born between 1905 and 1954.[8] Their results were similar to those reported by Mare (1981), who found no major changes in the logit coefficients for the impact of father's prestige on educational transitions among white males born between 1907 and 1951.

We restrict the GSS sample to respondents aged between 30 and 69 at interview and born in one of four birth cohorts: 1930-1939, 1940-1949, 1950-1959, or 1960-1969. The final sample consists of 16,077 individuals. To account for the sampling design, we apply the weight suggested by GSS (Smith *et al* 2011). Following Hout *et al* (1993) we report results for the pooled sample. Our dependent variable is highest level of educational attainment categorized as:

1 Less than high school

2 High school

[8] Hout et al. (1993) estimated family background effects for successive educational transitions, as suggested by Mare (1981). In addition to father's prestige, these models also included sex, parental education, and farm background.

3 Junior college

4 Bachelor

5 Graduate

We use father's socio-economic index (SEI) as our measure of social origins.[9] We fit an ordered logit model to each cohort, and, using equation (6), we derive the correlation between paternal SEI and the latent variable assumed to underlie the categorical measure of highest education. In Table 4 we show the trends in the coefficients from the ordered logit model and the corresponding, derived correlations. The logit coefficients indicate equalization in educational outcomes across cohorts born in the mid-20th century U.S and this trend is statistically significant (one-tailed test, five percent significance level).

-- TABLE 4 HERE --

But the trend in correlations runs in the opposite direction, suggesting increasing inequality, though, in fact, all the correlation coefficients are within an interval of 0.34 through 0.38 and we cannot reject a hypothesis of no change in them across cohorts. They suggest that SEI explains between 12 and 15 percent of the variance in the latent educational propensity over the entire period studied. Thus the substantive conclusions about trends in inequalities in educational attainment change, once we base our inferences on the correlation coefficient rather than the logit coefficient: the logit coefficients point to a decline in inequalities, while the correlation points to constancy. But, given equation (7), we can explain why this comes about. Table 4 shows that the dispersion of SEI increases dramatically across cohorts: its standard deviation is roughly 25% larger in the 1960-69 cohort than in the 1930-39 cohort. These changes in the

marginal distribution of SEI explain why logit coefficients show a modest trend towards equalization whereas the correlation shows growing inequality. The apparent equalization in the case of the US is very much an artifact of the changing distribution of social origins.

Conclusions

Although textbooks commonly relate the coefficients of non-linear probability models to the corresponding parameters of an underlying latent variable regression, a strong case can be made for interpreting them in terms of underlying correlations. The logit or probit coefficient is a transformation of the correlation between a predictor, *x*, and the latent *y**. The correlation coefficient is a standardized metric, invariant to differences in the marginal distributions of *x* and *y** across groups, and it may therefore be used in comparative social research, solving a part of the problem of making group comparisons using logits or probits identified by Allison (1999). Our solution will be appropriate only when it is useful to base comparisons on the correlation coefficient, and we sought to outline what those circumstances are and when they are likely to arise. Finally, the methods we have presented can readily be implemented using a Stata© program called *nlcorr* (http://ideas.repec.org/c/boc/bocode/s457289.html).

---

[9] This measure is already provided in the GSS data. It is based on the 1980 Census occupational scheme.

References

Achen, Christopher H. 1977. "Measuring Representation: Perils of the Correlation Coefficient." *American Journal of Political Science* 21:805-821.

Agresti, Alan. 2002. *Categorical Data Analysis*. New Jersey: Wiley.

Ai, Chunrong and Edward C. Norton. 2003. "Interaction terms in logit and probit models." *Economics Letters* 80:123-129.

Allison, Paul D. 1999. “Comparing Logit and Probit Coefficients Across Groups.” *Sociological Methods & Research* 28:186-208.

Björklund, Anders and Markus Jäntti 2009. ”Intergenerational Income Mobility and the Role of Family Background.” Pp. 491-521 in Wiemer Salverda, Brian Nolan, and Timothy M. Smeeding (eds) *The Oxford Handbook of Economic Inequality*. Oxford: Oxford University Press.

Blalock, Hubert M. 1964. *Causal Inference in Nonexperimental Research.* Chapel Hill: University of North Carolina Press.

Blalock, Hubert M. 1967. "Causal Inference, Closed Populations, and Measures of Association." *American Political Science Review* 61:130-136.

Breen, Richard and Jan O. Jonsson. 2005. "Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility." *Annual Review of Sociology* 31:223-243.

Breen, Richard and Kristian Bernt Karlson. 2013. "Counterfactual Causal Analysis and Nonlinear Probability Models." Pp. 167-188 in Stephen L. Morgan (ed.) *Handbook of Causal Analysis for Social Research*. New York: Springer.

Breen, Richard, Ruud Luijkx, Walter Müller, and Reinhard Pollak. 2009. “Nonpersistent Inequality in Educational Attainment: Evidence from Eight European Countries.” *The American Journal of Sociology* 114:1475-1521.

Breen, Richard, Ruud Luijkx, Walter Müller, and Reinhard Pollak. 2010. “Long-term trends in Educational Inequality in Europe: Class Inequalities and Gender Differences.” *European Sociological Review* 26:31-48.

Cox, N.R. 1974. “Estimation of the Correlation Between A Continuous and a Discrete Variable.” *Biometrics* 30:171-178.

Erikson, Robert and John H. Goldthorpe. 1992. *The Constant Flux. A Study of Class Mobility in Industrial Societies*. Oxford: Clarendon Press.

Greene, William. 2010. "Testing hypothesis about interaction terms in nonlinear models." *Economics Letters* 107:291-296.

Hargens, Lowell L. 1976. “A Note On Standardized Coefficients as Structural Parameters.” *Sociological Methods & Research* 5:247-256.

Hout, Michael, Adrian E. Raftery, and Eleanor O. Bell. 1993. “Making the Grade: Educational Stratification in the United States, 1925-1989.” Pp. 25-47 in Yossi Shavit and Hans-Peter Blossfeld (eds.) *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Boulder: Westview Press.

Karlson, Kristian Bernt, Anders Holm and Richard Breen 2012. “Comparing Regression Coefficients between Same-Sample Nested Models using Logit and Probit: A New Method” *Sociological Methodology* 42:286-313.

Kim, Jae-On and Charles W. Mueller. 1976. "Standardized and Unstandardized Coefficients in Causal Analysis: An Expository Note." *Sociological Methods and Research* 4:423-438.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30:666-687.

Linn, Robert L. and Charles E. Werts. 1969. “Assumptions in making causal inferences from part correlations, partial correlations, and partial regression coefficients.” *Psychological Bulletin* 72:307-310.

Long, J.S. 2009. “Group comparisons in logit and probit using predicted probabilities.” *Unpublished working paper (June 25, 2009).*

Mare, Robert D. 1981. “Change and Stability in Educational Stratification.” *American Sociological Review* 46:72-87.

McFadden, David. 1974. “Conditional Logit Analysis of Qualitative Choice Behavior.” Pp. 105-142 in P. Zarembka (ed.) *Frontiers in Econometrics.* New York: Academic Press.

McKelvey, Richard D. and William Zavoina. 1975. “A Statistical Model for the Analysis of Ordinal Level Dependent Variables.” *Journal of Mathematical Sociology* 4:103-120.

Olsson, Ulf, Fritz Drasgow and Neil J. Dorans. 1982. “The polyserial correlation coefficient.” *Psychometrika* 47:337-347.

Schoenberg, Ronald. 1972. "Strategies for Meaningful Comparison." Pp. 1-35 in Costner H.L. (ed.) *Sociological Methodology*. Los Angeles, CA: Sage.

Shavit, Yossi and Hans-Peter Blossfeld. 1993. *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Boulder: Westview Press.

Smith, Tom W., Peter V. Marsden, Michael Hout, and Jibum Kim. 2011. *General Social Surveys, 1972-2010. Cumulative Codebook*. Chicago: National Opinion Research Center.

Theil, Henri. 1972. *Statistical Decomposition Analysis*. Amsterdam: North-Holland Pub. Co.

Tukey, J.W. 1954. "Causation, regression, and path analysis." Pp. 35-66 in Oscar Kempthorne (ed.) *Statistics and Mathematics in Biology*. Ames: Iowa State College Press.

Williams, Richard. 2009. "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups." *Sociological Methods & Research* 37:531-559.

Appendices

Appendix A: Group Comparison under the Assumption of a Truly Discrete Variable

We show that—without assuming the existence of a latent outcome variable—estimated logit coefficients can differ between groups according to arbitrary heteroskedasticity in the binary outcome. The result follows from standard results on marginal versus conditional interpretations of nonlinear probability models, as discussed by Agresti (2002:498-501).

Write the logit for y = 1 for two groups,

$$\text{logit}(Y=1) = a_1 + b_1 x + \gamma_1 u_1 \quad \text{(A1)}$$

$$\text{logit}(Y=1) = a_2 + b_2 x + \gamma_2 u_2 , \quad \text{(A2)}$$

where u_j, j = 1,2, are unobserved variables assumed to be pair wise independent of x_j, and a_j are logit constant terms.

The role of the unobserved variable can differ between groups in two ways. First, the unobserved variables can have different effects, γ_j, on the binary outcome. Second, the distribution of unobservables, u_j, can differ. Because we cannot observe u_j, the unobservables are essentially "integrated out" of the estimating equation (i.e., the first derivatives of the log-likelihood function). In other words, we do not estimate parameters based on (A1) and (A2), but rather the averaged versions of the estimation equations, averaged over the distribution of the effect of the unobservables:

$$E(Y_j = 1) = E\left(\frac{\exp(a_j + b_j x + \gamma_j u_j)}{1 + \exp(a_j + b_j x + \gamma_j u_j)}\right); j = 1, 2. \quad \text{(A3)}$$

For general distributions of u_j, (A3) does not have a closed form solution. However, whenever $\gamma_j u_j$ are normally distributed, we have that

$$E(Y_j = 1) = \left(\frac{\exp\left(\delta_j\left(a_j + b_j x\right)\right)}{1 + \exp\left(\delta_j\left(a_j + b_j x\right)\right)}\right); j = 1, 2 \ ,$$

where $\delta_j = (1 + 0.6\sigma_j)^{-1/2}$, and where σ_j is the standard error of the distribution of $\gamma_j u_j$, with corresponding logit equations

$$\text{logit}(Y = 1) = \alpha_1 + \beta_1 x \quad \text{(A4)}$$

$$\text{logit}(Y = 1) = \alpha_2 + \beta_2 x \quad \text{(A5)}$$

where $\alpha_j = \delta_j a_j$, $\beta_j = \delta_j b_j$. The estimates based on (A4) and (A5) are usually referred to as marginal or population averaged parameters. Because b_j is multiplied by δ_j and because δ_j can differ across groups, the estimated population averaged parameters may differ across groups, even when $b_1 = b_2$. Because the variances of the unobservables σ_j are unobserved, the relation between the group-specific factors, δ_j, is not identified. In other words, group differences in the dispersion of unobservables, σ_j, might conflate any true difference in effects of x between groups. In effect, this result parallels that derived under the assumption of a latent, continuous outcome variables.

Because σ_j depend on both γ_j and u_j, we notice that the scale factors, δ_j, might differ when the unobservables have different effects across groups, different distributions across groups, or both. We further notice that this result does not pertain to linear models. In these models, $E(Y_j = y) = E\left(\alpha_j + \beta_j x + \gamma_j u_j + e_j\right) = \alpha_j + \beta_j x;\ j = 1,2$, when x and u are mean independent, i.e. $E\left(\gamma u + e \mid x\right) = 0$. Hence, in the linear model, the residual variance $\left(\text{var}\left(\gamma u + e\right)\right)$ and the mean are independent, so heteroscedasticity does not affect the conditional mean function.

Appendix B: Partial Correlation between x and y given z*

We show how to recover the partial correlation between *x* and the latent variable, *y**, controlling for one or more variables, *z*. Here we have:

$$\text{var}(y^*) = \beta^2_{y^*x.z}\ \text{var}(x) + \beta^2_{y^*z.x}\ \text{var}(z) + 2\beta_{y^*x.z}\beta_{y^*z.x}\ \text{cov}(x,z) + \text{var}(y^* \mid x,z)$$

with $\text{var}(y^* \mid x,z) = s^2\ \text{var}(\omega)$. The partial regression coefficient for *y** on *x* controlling for *z* is

$$\beta_{y^*x.z} = \frac{\text{cov}(y^*, x \mid z)}{\text{var}(x \mid z)},$$

and so we can write the partial correlation as

$$r_{y^*x.z} = \frac{\beta_{y^*x.z}\, sd(x \mid z)}{\sqrt{\text{var}(y^* \mid z)}} = \frac{b_{yx.z}\, s \cdot sd(x \mid z)}{\sqrt{\text{var}(y^* \mid z)}},$$

where *b* is the probit or logit coefficient. Using the result that

$$\text{var}(y^* \mid z) = \beta^2_{y^*x.z}\ \text{var}(x \mid z) + \text{var}(y^* \mid x,z),$$

we obtain (8).

Appendix C: Statistical Tests

Table A1 shows the results of a Monte Carlo analysis to study the power of the test given in equation (18). We simulated two populations, each of 5 million, one with $r(x,y^*)$=0.5, the other with a correlation varying from 0 to 0.45. We then drew a sample from each population and applied equation (18) to test the significance (using the .05 critical value) of the difference in their correlations. The figures in Table A1 show the proportion of times, out of 1000 replications, that the null hypothesis of no difference was rejected, using sample sizes ranging from 100 to 5000. In this case both x and y^* were normally distributed and the correlation was derived from a probit model. As we should have expected, the larger the difference in the correlation and the larger the sample size, the more likely the test is to reject the null hypothesis. With a sample of 1000 the test is effective at rejecting the null hypothesis when the difference in correlations is 0.2 or greater: for a sample of 5000 the test performs well for differences in correlations of more than about .07.

-- TABLE A1 HERE --

Appendix D: Coefficient of Determination

In some situations researchers are interested in summarizing the share of variation in y^* explained by the predictor variables. With a single predictor this is just the square of the correlation in (6). For simplicity, we derive the coefficient of determination assuming that we have two predictor variables, x and z.

The explained variance of the model in which y^* depends on both x and z is equal to $\text{var}(y^*) - \text{var}(y^* \mid x, z)$, where var(y*) is

$$\beta^2_{y^*x.z}\,\mathrm{var}(x) + \beta^2_{y^*z.x}\,\mathrm{var}(z) + 2\beta_{y^*x.z}\beta_{y^*z.x}\,\mathrm{cov}(x,z) + \mathrm{var}(y^* \mid x,z)$$

Because the coefficient of determination is the ratio of the explained to the total variance of y^* we have:

$$R^2 = \frac{\beta^2_{y^*x.z}\,\mathrm{var}(x) + \beta^2_{y^*z.x}\,\mathrm{var}(z) + 2\beta_{y^*x.z}\beta_{y^*z.x}\,\mathrm{cov}(x,z)}{\beta^2_{y^*x.z}\,\mathrm{var}(x) + \beta^2_{y^*z.x}\,\mathrm{var}(z) + 2\beta_{y^*x.z}\beta_{y^*z.x}\,\mathrm{cov}(x,z) + \mathrm{var}(y^* \mid x,z)}$$

Substituting the logit or probit coefficient and expanding $\mathrm{var}(y^* \mid x,z)$ we can rewrite the preceding equation as

$$R^2 = \frac{b^2_{yx.z}s^2\,\mathrm{var}(x) + b^2_{yz.x}s^2\,\mathrm{var}(z) + 2b_{yx.z}b_{yz.x}s^2\,\mathrm{cov}(x,z)}{b^2_{yx.z}s^2\,\mathrm{var}(x) + b^2_{yz.x}s^2\,\mathrm{var}(z) + 2b_{yx.z}b_{yz.x}s^2\,\mathrm{cov}(x,z) + s^2\,\mathrm{var}(\omega)}$$

The s^2 term cancels to leave a function of known values:

$$R^2 = \frac{b^2_{yx.z}\,\mathrm{var}(x) + b^2_{yz.x}\,\mathrm{var}(z) + 2b_{yx.z}b_{yz.x}\,\mathrm{cov}(x,z)}{b^2_{yx.z}\,\mathrm{var}(x) + b^2_{yz.x}\,\mathrm{var}(z) + 2b_{yx.z}b_{yz.x}\,\mathrm{cov}(x,z) + \mathrm{var}(\omega)}$$

This coefficient of determination equals that proposed by McKelvey and Zavoina (1975) and can easily be extended to the case with more than two predictor variables.

TABLES

TABLE 1. Monte Carlo study of sensitivity of correlation metric to pure misspecification of error term. 500 replications. Mean absolute deviation from true correlation reported.

		N = 200		N = 1,000		N = 5,000	
Scenario	True error distr.	Logit	Probit	Logit	Probit	Logit	Probit
r = 0.25							
1A	Normal	0.04	0.04	0.03	0.02	0.03	0.01
1B	Logistic	0.04	0.05	0.02	0.03	0.01	0.03
1C	t(6)	0.04	0.05	0.02	0.04	0.01	0.04
1D	Lognormal	0.04	0.05	0.02	0.03	0.01	0.02
1E	Uniform	0.08	0.06	0.09	0.06	0.08	0.06
r = 0.50							
2A	Normal	0.05	0.04	0.04	0.02	0.04	0.01
2B	Logistic	0.04	0.04	0.02	0.03	0.01	0.03
2C	t(6)	0.04	0.05	0.02	0.05	0.01	0.05
2D	Lognormal	0.04	0.05	0.02	0.03	0.01	0.03
2E	Uniform	0.12	0.08	0.12	0.08	0.12	0.08
r = 0.75							
3A	Normal	0.03	0.03	0.02	0.01	0.02	0.01
3B	Logistic	0.03	0.03	0.01	0.02	0.01	0.01
3C	t(6)	0.03	0.03	0.02	0.02	0.01	0.02
3D	Lognormal	0.03	0.03	0.01	0.02	0.01	0.02
3E	Uniform	0.06	0.04	0.06	0.03	0.06	0.03

Note: t(6) is the t distribution with six degrees of freedom.

TABLE 2. Monte Carlo study of power of statistical test of a single correlation. 1,000 replications. Proportion of rejections of H_0 reported.

Scenario	H_0	True corr.	N = 100	N = 200	N = 1,000	N = 2,000
A	False	0.00	5%	5%	4%	5%
B	True	0.10	16%	25%	82%	98%
C	True	0.25	61%	90%	100%	100%
D	True	0.50	100%	100%	100%	100%
E	True	0.75	100%	100%	100%	100%
F	True	0.90	100%	100%	100%	100%

TABLE 3. Decomposition of logit coefficients by country, cohort, and social class

	II		IIIa		IVab		IVc		V+VI		VII+IIIb	
Birth cohort	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$	$\frac{r}{\sqrt{1-r^2}}$	$\frac{sd(\omega)}{sd(x\mid z)}$
German men												
1908-24	-0.042	9.740	-0.137	9.945	-0.232	8.569	-0.380	8.557	-0.365	7.726	-0.453	8.286
1925-34	-0.099	9.879	-0.172	10.122	-0.236	9.517	-0.389	8.937	-0.347	8.199	-0.445	8.669
1935-44	-0.102	8.168	-0.164	8.985	-0.226	8.261	-0.361	8.495	-0.397	6.799	-0.459	7.414
1945-54	-0.080	8.026	-0.119	8.532	-0.150	8.746	-0.265	8.737	-0.329	6.586	-0.362	7.213
1955-64	-0.052	7.871	-0.095	8.605	-0.138	8.796	-0.249	9.113	-0.310	6.214	-0.355	6.952
French men												
1908-24	-0.060	13.616	-0.161	11.895	-0.225	9.840	-0.405	9.403	-0.276	9.879	-0.369	9.548
1925-34	-0.066	12.824	-0.168	11.299	-0.214	9.460	-0.433	8.976	-0.303	9.216	-0.390	9.024
1935-44	-0.107	11.419	-0.178	10.509	-0.245	8.547	-0.447	8.093	-0.341	8.128	-0.427	7.980
1945-54	-0.099	9.825	-0.195	9.746	-0.227	8.195	-0.357	7.936	-0.345	7.469	-0.419	7.492
1955-64	-0.085	9.313	-0.204	9.152	-0.197	8.375	-0.322	8.267	-0.332	7.007	-0.390	7.308
Dutch men												
1908-24	-0.051	9.311	-0.078	12.107	-0.202	7.840	-0.297	7.854	-0.245	7.843	-0.328	7.438
1925-34	-0.013	9.032	-0.055	11.286	-0.176	7.911	-0.257	7.932	-0.222	7.603	-0.307	7.337
1935-44	-0.038	8.094	-0.070	10.018	-0.167	7.474	-0.215	7.686	-0.246	7.043	-0.304	6.806
1945-54	-0.015	7.365	-0.055	9.098	-0.148	7.383	-0.153	7.782	-0.202	6.432	-0.239	6.489
1955-64	-0.007	6.703	-0.061	8.833	-0.088	7.459	-0.169	7.899	-0.194	5.800	-0.240	6.079

TABLE 4. Logit coefficients, correlations coefficients and standard deviations of SEI by US cohorts, GSS data 1988-2010

Birth cohort	Logit coef.	Correlation	Standard deviations of SEI
1930-1939	0.0424	0.3473	15.85
1940-1949	0.0400	0.3615	17.59
1950-1959	0.0368	0.3578	18.88
1960-1969	0.0377	0.3768	19.58

Note: Weight used (wtssall).

TABLE A1. Monte Carlo study of power of statistical test of the cross-sample difference between two correlations. Sample 1 is drawn from a population with a true correlation of 0.5. 1,000 replications. Proportion of rejections of H_0 reported.

Scenario	True corr. in sample 2	N = 100	N = 500	N = 1,000	N = 5,000
A	0.00	81%	100%	100%	100%
B	0.10	66%	100%	100%	100%
C	0.20	47%	98%	100%	100%
D	0.30	26%	79%	98%	100%
E	0.40	11%	33%	53%	100%
F	0.45	8%	13%	21%	63%

FIGURES

FIGURE 1. Ordered logit results

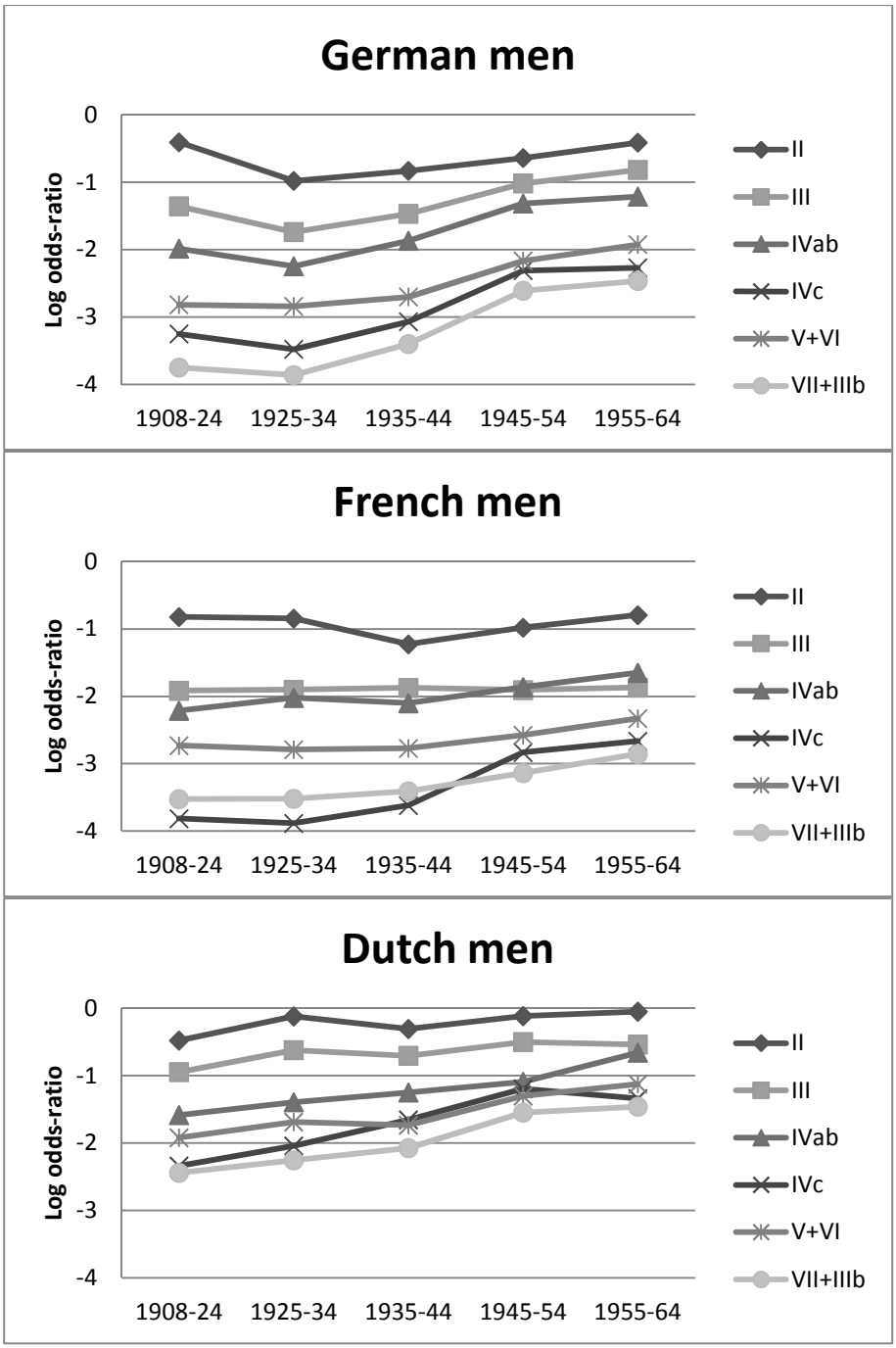

FIGURE 2. Partial correlation results

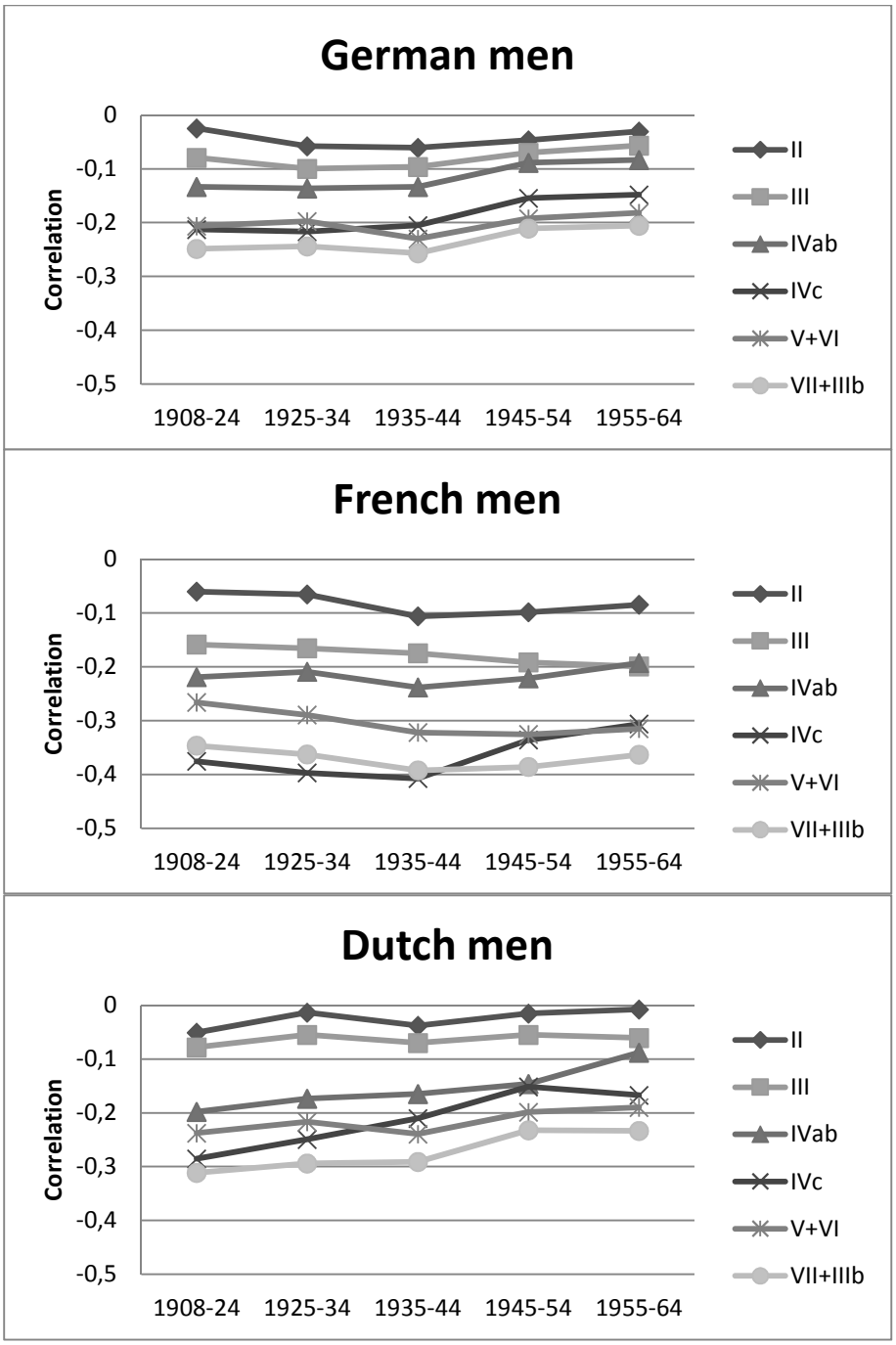

FIGURE 3. R-squared values by country and cohort